Wissenschaftstheoretische Anforderungen an empirische Forschung und die Problematik ihrer Beachtung in der Evaluation: oder: Wie sich die Evaluationsforschung um das Evaluieren drückt

Kromrey, Helmut

Wissenschaftstheoretische Anforderungen an empirische Forschung und die Problematik ihrer Beachtung in der Evaluation

Oder: Wie sich die Evaluationsforschung um das Evaluieren drückt[1]

Helmut Kromrey

Erste Vorbemerkung: Wo liegt das Problem?

Wenn von Evaluation – in welchem Zusammenhang und in welcher spezifischen Begriffsbedeutung auch immer – gesprochen wird, ist damit eine bewertende Aussage über einen Gegenstand oder Sachverhalt gemeint. Kommt dabei als Erkenntnisinstrument (auch) empirische Forschung ins Spiel, kann das Ziel von Evaluation präzisiert werden als: empirisch gestützte Gewinnung von Bewertungen mit intersubjektivem Geltungsanspruch. Oder zugespitzt auf das Thema dieser Veranstaltung: Es geht um die forschungsgestützte Gewinnung von *Qualitäts*aussagen. Das Ziel ist also eine normative Aussage, ein „Werturteil".

Hier liegt nun in der Tat ein gravierendes Problem; schließlich sind nach herrschender Meinung in Methodologie und Wissenschaftstheorie normative Aussagen – Werturteile generell und damit auch Qualitätsurteile im Besonderen – empirisch nicht begründbar. Umso mehr überrascht es, dass dies – abgesehen von ganz seltenen Ausnahmen[2] – bei Evaluationen trotz ihrer quantitativ zunehmenden Bedeutung als Problem nicht erkannt, zumindest jedenfalls nicht thematisiert wird, weder von Evaluationspraktikern noch von Methodologen derjenigen Schule empirischer Wissenschaft, die sich dem Wertneutralitätspostulat verpflichtet fühlt.

Die skizzierte Problematik wirft ein ganzes Spektrum von Fragen auf, die natürlich in einem einzigen Vortrag nicht abgehandelt werden können. Dies umso weniger als zugleich auch ein ganzes Spektrum unterschiedlicher Verständnisse von Evaluation ko-existiert, von relativ wenig formalisierter Evaluierung durch Experten (auch als Basis von professioneller Beratung) über responsive, stakeholder-orientierte, partizipative Evaluationsansätze bis hin zu unterschiedlichen forschungsbasierten Evaluationen oder auch „ganz gewöhnlichen" Evaluations-Befragungen. Ich beschränke mich hier auf die Wertproblematik im Rahmen von Evaluation durch Forschung (ungenauer: „Evaluationsforschung").

---

1 Überarbeitete Fassung eines Vortrags auf dem Kongress für Soziologie, Sitzung der Sektion Methoden empirischer Sozialforschung, am 11.10.2006 in Kassel: „Die empirische Erfassung von Qualität"

2 Beispielsweise Christian Lüders wie auch Wolfgang Beywl in Flick 2006.

Zweite Vorbemerkung: Wertproblematik und Evaluationsforschung – ein unlösbares Dilemma?

Im Rahmen analytisch-nomologischer Wissenschaftstheorie – eine weitere Einschränkung der thematischen Reichweite meines Vortrags – ist der Anspruch, „empirisch gestützte Bewertungen mit intersubjektivem Geltungsanspruch" zu gewinnen, methodologisch nicht ohne Weiteres legitimierbar:

- Die unmittelbare empirische Begründung von Bewertungen durch Forschung ist nicht möglich; auch aus korrekten *empirischen* Beschreibungen und Analysen sind *normative* Aussagen nicht ableitbar.
- Die Geltungsbegründung empirischer Analysen folgt einer anderen Logik als die Geltungsbegründung normativer Aussagen. Für die ersteren gibt es in der Wissenschaftstheorie klare Regeln, für die letzteren nicht.
- Möglich ist es lediglich, die Forschung möglichst genau auf den Zweck „Bewertung" auszurichten, indem wenn schon nicht direkt bewertende, so doch bewertungs*relevante* Informationen gesammelt und systematisiert werden.[3]

Soll also unmittelbar „durch Forschung" evaluiert werden, müssen Strategien gefunden werden, mit deren Hilfe erreichbar wird, dass die empirischen Daten einen quasi normativen Charakter erhalten, so dass sie sozusagen „für sich selbst sprechen". Eine explizite Geltungsbegründung daraus abgeleiteter Wertaussagen durch die Forschung wäre dann nicht mehr notwendig. Solche Strategien gibt es in der Tat; und drei von ihnen sollen hier skizziert werden.

Exkurs: Die Argumentationslogik analytisch-nomologischer, „wertfreier" Forschung

Für *wissenschaftstheoretische Argumentationen* kann das von Hempel und Oppenheim konzipierte Schema wissenschaftlicher Erklärung (Hempel/Oppenheim 1948) als verbindliches Gerüst gelten:

Explanans:	(1)	Es gilt (mindestens) ein nomologisches Gesetz (z.B.: „Wenn A und B, dann C")
	(2)	Die in der Wenn-Komponente genannten Randbedingungen sind empirisch erfüllt (z.B.: „A und B liegen vor")
Explanandum:	(3)	Singulärer Satz, der den zu erklärenden Sachverhalt beschreibt (z.B. „C liegt vor").

Gegeben ist das zu erklärende „singuläre Ereignis"(3), gesucht ist das „Explanans" (1 und 2). Bei dieser Art von Erklärung muss (3) deduktiv-logisch aus (1) und (2) folgen, wobei (2) aus der

3 Dies geschieht z.B. in der Hochschulevaluation und/oder bei Akkreditierungen nach dem bekannten mehrstufigen Modell des peer review. Die Forschung hat hier nur die Funktion des Informationszulieferers; das Evaluieren (d.h. das Fällen der Werturteile) geschieht durch Experten oder durch ein dazu legitimiertes Gremium oder durch Aushandeln zwischen den beteiligten Parteien.

Wenn-Komponente und (3) aus der Dann-Komponente des nomologischen Gesetzes abgeleitet wird.[4]

In der *alltäglichen empirischen Forschung*, die sich auf die analytisch-nomologische Wissenschaftstheorie beruft, scheint das H-O-Schema hingegen wenig praktische Bedeutung zu haben. Dieser Eindruck täuscht jedoch.[5]

Zum einen gilt die *Logik der Erklärung* auch für die Konstruktion des Forschungsdesigns zum empirischen *Test von Theorien und Hypothesen*. Lediglich das Erkenntnisinteresse ist ein anderes: Es sind nicht singuläre Ereignisse (3) „zu erklären", sondern es sind nomologische Hypothesen (1) auf ihre empirische Geltung „zu prüfen". Dazu werden empirische Testbedingungen entweder geschaffen (Experiment) oder in kontrollierter Weise aufgesucht und beschrieben (Survey), in denen die Randbedingungen (Wenn-Komponente) erfüllt sind (2). Und bei Vorliegen dieser Randbedingungen wird geprüft, ob auch die in der Dann-Komponente behaupteten Konsequenzen auftreten bzw. aufgetreten sind (3).

Für dieses Design dürfte die skizzierte Logik auch von Forschungspraktikern leicht einzusehen sein. In anderen Forschungszusammenhängen dagegen werden seit Poppers Fokussierung der Methodologie auf die Testlogik wissenschaftstheoretische Grundlagen eher stiefmütterlich behandelt.

Bei genauerem Hinsehen aber können die drei Komponenten des H-O-Schemas z.B. auch im deskriptiven Survey-Modell - also bei empirischen Forschungen zum Zwecke deskriptiver Diagnose sozialer Problemfelder - wieder gefunden werden. Als methodologisches Gerüst ist in diesem Design nach Formulierung einer präzisen Fragestellung ein deskriptives Modell des zu untersuchenden Gegenstands auszuarbeiten. Dieses wird idealtypischerweise auf der Basis „empirisch bewährter" Theorien (1) entwickelt, im Realfall faktischer Forschung ergänzt um Hypothesen möglichst hoher Plausibilität. Dieses Modell hat eine forschungsleitende Funktion, dient sozusagen als „Wegweiser" im Forschungsprozess, ist also Erkenntnis*basis* und nicht Gegenstand der Überprüfung. Da jedoch die erhobenen Daten der empirischen „Diagnose" sozialer Probleme dienen sollen, müssen sie „Erklärungswert" haben und sowohl die Problemdimensionen differenziert beschreiben (3) als auch die relevanten Randbedingungen (2) für das Auftreten der Probleme erfassen.

Das H-O-Erklärungsschema ebenso wie die Übernahme seiner Struktur für andere Erkenntniszwecke impliziert als *erkenntnistheoretische* Basis den erkenntnistheoretischen Realismus. D.h., unterstellt wird auf der Gegenstandsseite eine „real existierende" Welt, gekennzeichnet durch Merkmale wie Ordnung, Struktur und Tatsachenautonomie, die Geltung von Regelhaftigkeiten bzw. Gesetzmäßigkeiten und Kausalität. Unterstellt wird zudem auf der *Seite des erkennenden Subjekts* die prinzipielle, wenn auch möglicherweise unvollständige und teilweise fehlerbehaftete Erkennbarkeit dieser Realität durch Wahrnehmungssinne sowie unterstützende Instrumente.

Das *Ziel* empirischer Wissenschaft ist hier die Erkenntnis der „wahren" Strukturen und Gesetzmäßigkeiten der Realität sowie ihre Dokumentation in Theorien. Erreicht werden soll dies durch eine *Strategie* des „kontrollierten Ratens" über das Aufstellen erkenntnisleitender ex-ante-Hypothesen und deren Konfrontation mit der (*objektiven*) Realität, abgebildet in (*subjektiven*) Wahrnehmungsda-

4 Analog ist der Argumentationszusammenhang bei einer Je-desto-Beziehung. Das „Gesetz" (1) könnte lauten: „Je höher X, desto höher Y". Empirisch erfüllt sein müsste die Randbedingung (2): „X ist gestiegen". Erklärt wäre damit das Explanandum (3): „Y ist gestiegen".

5 Ausführlicher dazu Kromrey 2006, S. 87 ff.

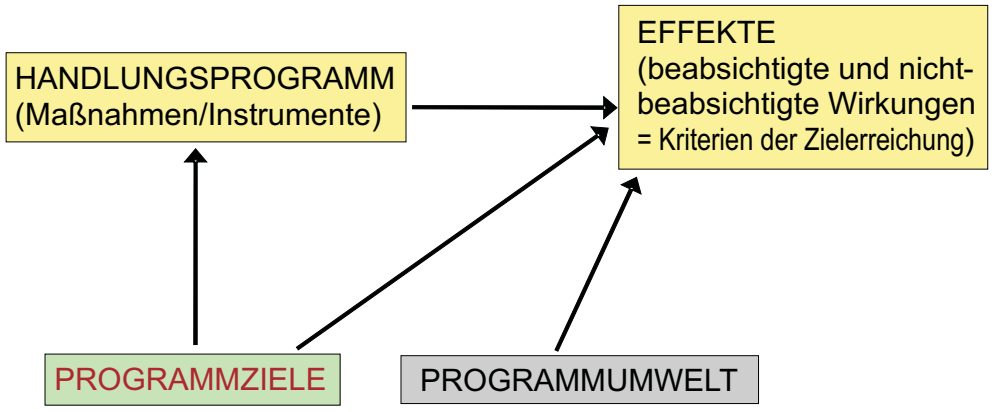
ten. Die darin implizierte Subjektivität wird kontrolliert durch strenge methodologische Regeln, die ein möglichst weitgehendes Ausschalten wahrnehmungsbeeinflussender Bedingungen bei der empirischen Erkenntnisgewinnung gewährleisten sollten („Objektivierung" der Verfahren, intersubjektive Nachprüfbarkeit).

Eine besondere Bedeutung kommt hier der Trennung deskriptiver (und damit „objektivierbarer") von normativen Aussagen zu, deren subjektiver Charakter methodologisch nicht aufhebbar und deren intersubjektive Geltung daher mit empirischen Mitteln nicht begründbar ist. Die normativen Elemente der Erkenntnisgewinnung werden daher aufgeteilt in solche, die die *normative* Basis der Forschung bilden (der Forschung *vor*gelagerte *wissenschaftsimmanente* Werte), und den *nichtwissenschaftlichen* Interessen und Werten, die aus dem wissenschaftlichen Begründungskontext *aus*gelagert und in den Entstehungs- und Verwertungskontext des Forschungsprojekts verwiesen werden.

In genau dieser pragmatischen Strategie zur Handhabung des Werturteilsproblems wird in der „herkömmlichen empirischen Forschung" auch die Lösung des Bewertungsdilemmas in der Evaluationsforschung gesucht.

„Wertneutrale Evaluation" im Design der Programmevaluation: der methodologische Idealtypus

Die oben skizzierte Forschungslogik einer auf Wertfreiheit verpflichteten Wissenschaft ist bruchlos auf die Evaluationsforschung anwendbar, sofern es sich bei ihrem Gegenstand um ein ausgearbeitetes „Programm"[6] handelt, das in Form explizit ausformulierter Programmziele die für eine „empirische Bewertung" notwendige normative Basis bereits in die Evaluation mitbringt. Das Design dieses Evaluationstyps berücksichtigt drei Dimensionen des zu bewertenden Gegenstands - Ziele, Maßnahmen, Effekte - sowie die programmexterne Umwelt als Quelle möglicher „Störvariablen":



6 Programme sind komplexe Handlungsmodelle, die auf die Erreichung expliziter Ziele gerichtet sind, die auf bestimmten, den Zielen angemessen erscheinenden Handlungsstrategien beruhen, und für deren Abwicklung finanzielle, personelle und sonstige Ressourcen bereitgestellt werden (Hellstern/Wollmann 1983, 7; grundlegend: Mayntz 1980).

Auch dieses Konzept verwendet die *Hempel-Oppenheimsche* Logik. Zum einen hat schon das „*Programm*“ als „technologische Aussage“ formal die gleiche Struktur wie eine „Erklärung“:

- Programmziele = angestrebte künftige Situation = Punkt (3) im *H-O-Schema*;
- „Maßnahmen“ = zu vollziehende Eingriffe in die gegenwärtigen „Randbedingungen“ = Punkt (2);
- die Art und Weise, *wie* eingegriffen werden soll, beruht auf Annahmen über Ursache-Wirkungs-Prinzipien = theoretische Basis = Punkt (1).

Auch das Design der Evaluation orientiert sich an diesen Komponenten:

- Sowohl die existierenden Randbedingungen (2) als auch der Ist-Zustand der Zielvariablen (3) sind vor Programmbeginn – t_0 – empirisch zu beschreiben.
- Während der Programmlaufzeit sind die Veränderungen der Randbedingungen (2) zu erfassen („monitoring“ sowohl der Eingriffe durch die im Programm vorgesehenen Maßnahmen als auch anderer relevanter Veränderungen in der Programmumwelt).
- Schließlich ist sicherzustellen, dass der Zustand der Zielvariablen (3) nach Programmdurchführung – t_1 – wiederum empirisch beschrieben wird, so dass Art und Ausmaß der Veränderungen feststellbar sind.

Der Gegenstand, den es zu bewerten gilt, ist natürlich als Gegenstand nicht wertneutral oder zweckfrei. Ganz im Gegenteil: Das Programm soll etwas erreichen. Damit wird auch das Konzept „Evaluation als Programmwirkungsforschung“ mit dem Wertproblem konfrontiert. Für die Forschung ist es aber dadurch gelöst worden, dass es in den „Entstehungskontext“ verlagert wurde (Programmziele als normative Basis), wodurch die eigentliche „Evaluation“ einen deskriptiven (und somit „wertneutralen“) Charakter erhält:

Der Bewertungsprozess reduziert sich hier auf einen Vergleich der vom Programm gesetzten Sollwerte (Zielerreichungskriterien) mit den gemessenen (und den Maßnahmen zurechenbaren) Effekten im Wirkungsfeld des Programms. Solche Aussagen lassen sich in vollem Umfang auf die Logik der Geltungsbegründung empirischer Faktenbehauptungen stützen. Sofern von der Evaluation differenzierte „*Qualität*surteile“ gefordert werden, bestehen diese in Aussagen über die Zielerreichung, ggf. konkretisiert durch die Dimensionen Effektivität (Wirkungsgrad der jeweiligen Maßnahmen) und Effizienz (Kosten-Wirksamkeit-Relation).

Dass die Realisierung dieser Aufgaben riesige Schwierigkeiten bereitet, tut der überzeugenden *Logik* des Modells keinen Abbruch.

Anders fällt das Urteil aus, wenn es um die praktische Bedeutung für ein breites Anwendungsspektrum und für den „alltäglichen“ Evaluationsbedarf geht. Die methodologischen Schwierigkeiten sind nämlich leider so riesig, dass diese überzeugende Logik nur unter sehr einschränkenden, nur ganz selten erfüllbaren Bedingungen praktisch einsetzbar ist. Überwiegend müssen wir uns mit mehr oder weniger guten Annäherungen an das idealtypische Modell zufrieden geben: z.B. in Form quasi-experimenteller Designs mit relativ vielen Zugeständnissen an die interne Validität, projektbegleitendem Monitoring mit erst in der Auswertungsphase an die Experimentallogik angelehnter statistischer Analyse, mit Zeitreihenbetrachtungen der Zielvariablen u.ä. Es liegt also nahe, nach komplett andersartigen Ersatzlösungen zu suchen.

Eine solche Ersatzlösung folgt aus der Überlegung, „Qualität" nicht erst anhand von „Effekten", also als *Folge* der Eigenschaften des zu bewertenden Gegenstands oder Sachverhalts zu interpretieren, sondern *unmittelbar* zu messen.

Evaluation als Qualitätsmessung: das methodologische Problemkind

Wenn es gelänge, am zu evaluierenden Gegenstand oder Sachverhalt Qualitätsmerkmale zu bestimmen und präzise zu definieren, hätten wir einen direkten Weg, die Evaluation methodologisch zu „okjektivieren", d.h. am zu bewertenden „Objekt" festzumachen. Wir könnten uns damit zugleich den schwierigen Umweg über die Messung von Outcome-Variablen und die Zurechnung ihrer Veränderungen als Zielerreichungs-Kriterien ersparen. Die implizite und unmittelbar einleuchtend scheinende Annahme bei dieser Überlegung ist: Wenn die Qualität des zu bewertenden Sachverhalts hoch ist, dann werden auch seine Wirkungen positiv sein; oder anders formuliert: dann werden auch die mit ihm verknüpften Ziele erreicht werden.

In diesem Fall hätte die Forschung lediglich die – forschungsmethodisch zur Alltagsroutine zählende – Aufgabe zu erfüllen, „Qualität" durch einen Satz qualitätsrelevanter Merkmale auszudifferenzieren und durch geeignete Indikatoren so zu operationalisieren, dass an ihnen situationsunabhängige „Qualitätsmesswerte" abgelesen werden können. Auch bei dieser Strategie wird die normative Basis der Evaluation in den Entstehungskontext ausgelagert, in dem der normative Begriff „Qualität" von dazu legitimierter Seite festzulegen ist.

Methodologisch ist diese Aufgabe allerdings gar nicht so simpel, wie es auf den ersten Blick erscheinen mag. Die „Sozialindikatorenbewegung" in den 1970er Jahren hat sich damit intensiv auseinandergesetzt und eine Liste von Anforderungen an Indikatorensysteme formuliert, [7] die auch in unserem Fall Geltung beanspruchen kann.

a) Die einzubeziehenden Indikatoren müssen in einem sachbezogenen Zusammenhang zu dem Problembereich stehen, für den Aussagen getroffen werden sollen (Gültigkeit).

b) Durch vorherige Aufstellung von Hypothesen ist zu gewährleisten, dass die Zusammenhänge zwischen dem einzelnen Indikator und anderen Indikatoren bzw. dem Indikatorsystem sichtbar werden. Im Idealfall wären die Indikatoren aus einem theoretischen Wirkungsmodell abzuleiten. (Theoriebezug).

c) Es ist zu begründen, warum eine bestimmte Maßzahl zur Präsentation ausgewählt wurde (Offenlegung des Wertbezugs).

d) Vor allem solche Indikatoren sind heranzuziehen, die für Gegenwart und Zukunft – möglichst auch für die Vergangenheit – erreichbar sind (Zeitbezug, Dauerhaftigkeit).

e) Die Indikatoren sollten sich auf solche Sachverhalte beziehen lassen, die durch politische Maßnahmen beeinflussbar sind („policy variables"; Praxisbezug).

Grundlegender ist jedoch eine andere Problematik, die in dieser Argumentation häufig übersehen wird: Neben dem Zutreffen der o.g. impliziten Annahme eines direkten Zusammenhangs zwischen Qualitätsmerkmalen und Zielerreichung (was empirisch geprüft werden kann) muss nämlich eine

7 s. z.B. Werner 1975.

weitere, eine erkenntnistheoretische (und damit *axiomatische*) Voraussetzung erfüllt sein: Qualität muss als direkte Eigenschaft des Objekts verstanden werden können (wie etwa Größe, Gewicht, Farbe usw.); bzw. methodologisch formuliert: Das Konstrukt „Qualität" ist so zu definieren, dass seine Dimensionen als Merkmale des Gegenstands erscheinen. Schon eine oberflächliche semantische Analyse lässt erkennen, dass „Qualität" eben *nicht* als Merkmal des zu bewertenden Objekts selbst zu verstehen ist, sondern als ein *relationales* Merkmal: als Eignung, Brauchbarkeit, Güte *in Bezug auf* bestimmte Ziele und Zwecke sowie auf bestimmte Nutzer- und Klientengruppen.[8]

Damit (und wenn wir die methodologischen Anforderungen der Sozialindikatorenbewegung einbeziehen) haben wir aber wieder das Anforderungsniveau des Modells der Programmevaluation erreicht - ohne allerdings über deren methodologische Problemlösung zu verfügen, nämlich: vom Programm vordefinierte Ziele und Maßnahmen. Anders als zu Beginn postuliert, bleibt - wenn sich das Konstrukt „Qualität" nicht aus dem Objekt selbst herleiten lässt - damit die Wert-Entscheidungs-Frage unbeantwortet. Die normative Basis für die Bewertung durch Qualitätsindikatoren setzt (und das heißt: die eigentliche Evaluation betreibt) diejenige Instanz, die festlegt, was als „Qualität" gelten soll und welche Qualitätskriterien und –standards anzulegen sind. Wenn aber diese Instanz die Nebenziele verfolgt, a) das ganze Evaluationsverfahren solle möglichst einfach und ohne großen Ressourceneinsatz zu bewerkstelligen sein und b) die Resultate dürften nicht noch interpretationsbedürftig, sondern sollten selbsterklärend sein (d.h.: höherer Messwert = „mehr Qualität"), dann passiert das, was wir vermutlich alle aus unseren eigenen Arbeitsumwelten kennen:

- In den Rang von Qualitätsindikatoren wird das erhoben, worüber Daten zur Verfügung stehen.
- Herangezogen werden vor allem quantitativ messbare Merkmale, so dass vorzugsweise Qualität durch Quantität operationalisiert wird.

Zusammengefasst: *Qualität ist, was leicht messbar ist*.

Angesichts dieser erneuten Problematik verwundert es nicht, wenn die Forschung sich aus diesen Dilemmata zu befreien versucht, indem sie Evaluation auf das reduziert, was sie unbestritten kann: Befragungen durchführen.

Evaluation durch Befragung: das erkenntnistheoretische Problemkind

Statt aufwändige, methodisch kontrollierte Evaluation durch Programmforschung zu betreiben (deren Anwendungsvoraussetzungen selten erfüllbar sind) oder Qualitätsindikatoren zu messen (deren Gültigkeit fragwürdig ist), wird die Bewertung von Maßnahmen (oder Sachverhalten oder Dienstleistungen) per „Betroffenenbefragung" ermittelt.

Die Adressaten und Nutzer, die Kunden und Klienten sind – so wird argumentiert – die von den zu evaluierenden Leistungen ganz konkret „Betroffenen" und daher in der Lage, aus eigener Erfahrung auch deren Qualität sachverständig und zuverlässig zu beurteilen. Befragungen erscheinen erheblich weniger anspruchsvoll – sowohl hinsichtlich des Aufwands der Durchführung als auch hinsichtlich

8 Man spiele als Gedankenexperiment einmal die vergleichsweise sehr simple Aufgabe durch, die „Qualität von Autoreifen" durch einen Satz situationsunabhängig geltender Merkmale messbar zu machen, also ohne Berücksichtigung von Einsatzkontexten (wie Stadtverkehr, Autobahn, Landstraßen, sportliche Wettbewerbe: Rallye Paris-Dakar oder Formel I) und ohne Berücksichtigung von Nutzergruppen (wie „Normalfahrer", LKW-Fahrer, Rennsportler).

der Strategie der Objektivierung: Sind die erbrachten Dienstleistungen „schlecht", so werden auch die Beurteilungen auf einer vorgegebenen Skala negativ ausfallen und umgekehrt. Befragt man eine hinreichend große Zahl von „Betroffenen" und berechnet pro Skala statistische Kennziffern (etwa Mittelwerte oder Prozentanteile), dann kommen – so die weitere Argumentation – individuelle Abweichungen der einzelnen Urteilenden darin nicht mehr zur Geltung.

So hieß es z.B. in einem Dozentenkurs des HDZ Essen schon 1980 optimistisch zum Thema Lehrevaluation:

- „Urteile (Schätzungen) von Studenten über die Lehre ... sind - wenn man etwa 20-30 Studenten urteilen lässt - zuverlässig wie professionelle Testverfahren".
  Und: „Sie sind von anderen Merkmalen der Studenten selbst und der Dozenten wenig beeinflusst." (Schmidt 1980, 51 f.)

Sofern dies zuträfe,[9] wäre die Evaluation per Befragung *der* Königsweg zur Lösung aller Probleme der Evaluationsforschung - auch der Werturteilsproblematik, denn die Bewertungen nehmen hier die „per Betroffenheit dazu Legitimierten" vor. Die Forschung selbst bleibt neutral; sie erhebt, systematisiert und analysiert lediglich.

Zwar sind „Messungen" per Befragung nicht so problemlos wie dies dem Laien häufig erscheint (s. Kromrey 2006, S. 257 ff.). Doch sofern systematische Verzerrungen vermieden werden können, lassen sich in der Tat bei hinreichend großer Befragtenzahl und bei repräsentativer Datenbasis individuelle Unterschiede - wie im obigen Zitat behauptet - „herausmitteln". Einzulösen sind hierfür lediglich durch das Erhebungsinstrument und in der Befragungssituation einige formale methodologische Voraussetzungen - die allerdings nicht ohne weiteres als erfüllt gelten können:

- der „Gegenstand" (das Objekt) der Beurteilung ist eindeutig definiert,
- das zu messende „Merkmal" ist eindeutig definiert und operationalisiert,
- eine „Mess-Skala" existiert und ist eindeutig definiert,
- die Befragten sind in der Lage, den „Gegenstand" intersubjektiv übereinstimmend zu identifizieren, das zu messende „Merkmal" intersubjektiv übereinstimmend zu erkennen und die „Mess-Skala" in intersubjektiv übereinstimmender Weise anzuwenden.

Im Falle der Erhebung von Evaluationen wird die Situation zusätzlich dadurch schwieriger, dass es sich bei den zu messenden Merkmalen um die oben genannten „qualitätsrelevanten Merkmale" (oder „Qualitätskriterien") handelt, durch die der Begriff „Qualität" operationalisiert wird. Und die hierauf anzuwendende Mess-Skala ist die Bezugsgröße, auf der das „Ausmaß" von Qualität angebbar ist (also der „Qualitätsstandard"). Damit sind wir aber auch bei der *indirekten* Qualitätsmessung per Befragung wieder mit dem gleichen Problem konfrontiert wie beim Ansatz der *direkten* Messung von Objektqualität durch Indikatoren.

Das statistische „Ausmitteln" von Messungenauigkeiten setzt bekanntlich die Existenz eines „wahren Wertes" voraus, von dem die einzelnen Messwerte lediglich „zufällig" abweichen. Bezogen auf die Beurteilungsvariation zwischen den einzelnen Befragten heißt dies: Um auf diese Weise zu einem gültigen Qualitätsmaß zu kommen, muss die Annahme gerechtfertigt sein, dass es einen „wahren" Qualitätswert für den zu beurteilenden Sachverhalt gibt, um den die einzelnen Antworten „zu-

9 Leider ist dieser Optimismus bei Lehrevaluationen nicht gerechtfertigt, wie differenzierte statistische Analysen von Daten aus Veranstaltungsbefragungen belegen (s. z.B. Kromrey 1994, 1995).

fällig“ streuen. Diese Annahme wäre aber nur dann haltbar, wenn eines der beiden folgenden erkenntnistheoretischen Axiome zuträfe:

- **Alternative 1:** Qualität ist ein „objektives“ Merkmal eines Sachverhalts,[10] dessen Ausprägung durch abbildende subjektive Wahrnehmung ohne systematische Verzerrung „gemessen“ werden kann, so dass bei hinreichend großer Zahl von Messungen der Erwartungswert dem „*wahren objektiven Wert*“ entspricht. 0der:
- **Alternative 2:** Qualität ist ein „intersubjektiv gültiges“ Konzept, über das alle Menschen in gleicher Weise verfügen. Anders formuliert: Alle Menschen bewerten nach gleichen Kriterien und Standards in gleicher Weise. In konkreten Situationen auftretende Unterschiede zwischen Bewertern sind als Zufallsvariation anzusehen, so dass bei hinreichend großer Zahl von Messungen der Erwartungswert dem „*wahren subjektiven Wert*“ entspricht.

Grundlage für die erste Alternative ist die im erkenntnistheoretischen Realismus (vom frühen Empirismus bis zum Gründer des Wiener Kreises, Moritz Schlick) vertretene Überzeugung von der Möglichkeit abbildender Wahrnehmung der Realität: Das Wahrgenommene steht in einem genauen Entsprechungsverhältnis zum Wirklichen. In diesem Fall wäre „Qualität“ durch standardisierte Befragung „objektiv“ messbar. Allerdings könnte sie dann auch durch direkte Indikatorenmessung (s.o.) „objektiviert“ werden.

Grundlage für die zweite Alternative wäre der erkenntnistheoretische Idealismus, am kompromisslosesten konzipiert in Platons „Ideenlehre“: Hinter der sinnlich wahrnehmbaren Welt stehen (als das „in Wahrheit Seiende“) die „Ideen“, die zwar der direkten Wahrnehmung nicht zugänglich, aber der unsterblichen Seele des Menschen von Anfang an mitgegeben sind. Erkenntnis besteht nach dieser Vorstellung im *Wiedererkennen* der allgemeingültigen Konzepte (der „Ideen“) in den empirischen (Einzel-)„Erscheinungen“ (s. Beckmann 1994, H. 3). Sofern „Qualität“ der Status einer solchen „Idee“ zukäme (analog zu Gerechtigkeit, Gleichheit, Heldentum, Liebe etc.), wäre sie durch standardisierte Befragung „intersubjektiv“ messbar.

Fazit

Wenn die Versuche wenig überzeugend erscheinen, das Wertproblem der Evaluationsforschung dadurch zu entschärfen, dass man die für das „wissenschaftliche Evaluieren“ erforderliche Wertbasis aus dem Begründungskontext empirischer Forschung hinausverlagert (in den Entstehungskontext oder in die extern vorzunehmende Definition eines normativen Begriffs von „Qualität“ oder in nicht intersubjektiv nachvollziehbare Bewertungsprozesse von „Betroffenen“), dann bieten sich aus meiner Sicht nur zwei Alternativen für eine „wissenschaftliche Evaluation“ an.

Die eine besteht darin, die Evaluation als einen Spezialfall aus dem Aufgabengebiet einer wertneutral verfahrenden empirischen Forschung auszusondern und ihr die zusätzliche Aufgabe der (nach wissenschaftlicher Methodologie verfahrenden, intersubjektiv nachprüfbaren) Ableitung von Wertaussagen zuzuschreiben. Überlegungen in dieser Richtung werden von Christian Lüders (2006) angestellt. Eine überzeugende Methodologie ist allerdings für mich derzeit nicht erkennbar.

10 Dies ist die identische Voraussetzung, die auch im Konzept der direkten Qualitätsmessung durch Indikatoren erfüllt sein muss.

Die andere Alternative besteht darin - und diese halte ich für die angemessenere Variante -, Evaluieren und Forschen klar zu trennen. Der Forschung ist die Aufgabe zuzuschreiben, alle für die Bewertung von Programmen, Maßnahmen etc. relevanten Informationen unter Einsatz des bewährten empirischen Instrumentariums zu erheben, zu analysieren und für Bewertungs- und Entscheidungsprozesse aufzubereiten. Die Funktion des Evaluierens sowie der Ableitung möglicher Konsequenzen für das Evaluationsobjekt sollte dagegen einem dafür explizit legitimierten Gremium zugewiesen werden. Selbstverständlich kann (und sollte) auch in diesem Prozess die Forschung beratend mitwirken. Dass dieses Modell realisierbar ist und die Akzeptanz von Evaluation erhöht, zeigt das bereits angesprochene Modell der mehrstufigen Hochschulevaluation, beispielsweise in der vom Verbund Norddeutscher Universitäten praktizierten Variante: Selbstbeschreibung / Selbstevaluation – peer review - Auswertende Konferenz (Nordverbund 2004).

Literatur

Beckmann, J. P., 1994: Einführung in die Erkenntnistheorie, Hagen: Fernuniversität Hagen (Kurs 3303)

Beywl, W., 2006: Evaluationsmodelle und qualitative Methoden. In: U. Flick 2006, 92-116

Flick, U.(Hg.), 2006: Qualittive Evaluationsforschung. Konzepte, Methoden, Umsetzungen. Reinbek bei Hamburg: Rowohlt

Hellstern, G.-M.; Wollmann, H., 1983: Evaluierungsforschung. Ansätze und Methoden, dargestellt am Beispiel des Städtebaus, Basel, Stuttgart

Hempel, C. G.; Oppenheim, P. C., 1948: Studies in the Logic of Explanation; in: Philosophy and Science, Vol. 15, 135-175

Kromrey, H., 1994: Wie erkennt man „gute Lehre“? Was studentische Vorlesungsbefragungen (nicht) aussagen. In: Empirische Pädagogik, 1994/2, 153-168

Kromrey, H., 1995: Evaluation der Lehre durch Umfrageforschung? Methodische Fallstricke bei der Messung von Lehrqualität durch Befragung von Vorlesungsteilnehmern. In: P. Mohler (Hg.): Universität und Lehre. Ihre Evaluation als Herausforderung an die Empirische Sozialforschung, Münster: Waxmann, 105-128

Kromrey, H., 2006: Empirische Sozialforschung. Modelle und Methoden der standardisierten Datenerhebung und Datenauswertung, 11. Aufl., Stuttgart: Lucius&Lucius, utb 1040

Lüders, Ch., 2006: Qualitative Evaluationsforschung – Was heißt hier Forschung? In: U. Flick 2006, 33-62

Mayntz, R., 1980: Die Entwicklung des analytischen Paradigmas der Implementationsforschung. In: dies. (Hg.): Implementation politischer Programme, Königstein/Ts., 1-17

Nordverbund (Hg.), 2004: 10 Jahre Evaluation von Studium und Lehre. Verbund Norddeutscher Universitäten. Verbund Materialien Band 16, Hamburg

Schlick, M., 1918: Allgemeine Erkenntnislehre (2. Aufl. 1925), unveränd. Nachdruck Frankfurt: Suhrkamp, 1978

Schmidt, J., 1980: Evaluation. I. Evaluation als Diagnose, Essen:HDZ

Werner, R., 1975: Soziale Indikatoren und politische Planung. Einführung in Anwendungen der Makrosoziologie, Reinbek: rororo;

Zur Person

Prof. Dr. Helmut Kromrey, Universitätsprofessor i.R. der Freien Universität Berlin, Empirische Sozialforschung, und Adjunct Professor of Sociology der Graduate School of Management der Universität Educatis (Altdorf/Schweiz).

Kontakt: mail@hkromrey.de